1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1992	

4. TITLE AND SUBTITLE
FROM COMPARISON DENSITY TO TWO SAMPLE ANALYSIS

5. FUNDING NUMBERS
DAAL03-90-6-0069

6. AUTHOR(S)
EMANUEL PARZEN

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Texas A&M University
Department of Statistics
College Station, Texas 77843-3143

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARO 27574.8-MA

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)* Contents of the paper are: 1. Introduction. 2. Quantile domain functions: F,Q; 3. Mid distribution and quantile functions: F^{mid}, F^{mid}; 4. Sample distribution and quantile: $F^{\sim}$, $Q^{\sim}$; 5. Comparison distribution and comparison density for discrete F,G; 7. Information measures: Renyi, Chi-square; 8. Information for comparison density function; 9. Information measures and entropy tests of fit; 10. Continuous versions of discrete distribution functions: F^{c}, Q^{c}; 11. Comparison distributions of one sample (continuous data); 12. One sample parameter estimation; 13. Comparision distributions of two samples (continuous) data; 14. Two sample comparison density estimation.

92 8 0

14. SUBJECT TERMS	15. NUMBER OF PAGES
statistical modeling, information measures, entropy, quantile functions, density estimation, quantile density estimate, comparison density, maximum spacings parameter estimation, minimum information parameter esitmation, goodness of fit.	20
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

FROM COMPARISON DENSITY TO TWO SAMPLE ANALYLSIS

Technical Report # 169

May 1992

Emanuel Parzen

Texas A&M Research Foundation

Project No. 5641

Sponsored by the U. S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

FROM COMPARISON DENSITY TO TWO SAMPLE ANALYSIS

Emanuel Parzen
Department of Statistics
Texas A&M University
College Station, Texas 77843-3143 USA

1. Introduction

This paper on statistical data modeling is written to express my esteem for Professor Hirotugu Akaike in celebration of his 65th birthday by a U.S./Japan Conference on *The Frontiers of Statistical Modeling: An Informational Approach* (held May, 1992 at the University of Tennessee). I like to play the game "how long have you known Professor Akaike" because I have the good fortune of knowing him since 1965.

Parzen (1979) argues that statistical data analysis should be defined as fitting probability models to data. This paper presents typical concepts and recent results of our modeling theory which emphasizes quantile domain functions, information measures, and comparison density estimation. Ultimate goals include: unify parametric and nonparametric inference for continuous and discrete data; demonstrate that mathematical statistical and data analytic approaches are both needed for statistical inference; stimulate exoteric methods (applicable by applied researchers) rather than esoteric methods (known only to a small group of mathematical statisticians); combine mathematical statistical and data analytic views to develop methods of statistical analysis which are based on assumptions (known model) which are tested in ways that provide insight how to model deviations of the data from the assumed model (and thus identify a "true" model as an "iterated" model).

Contents of the paper are: 1. Introduction; 2. Quantile domain functions; F, Q; 3. Mid distribution and quantile functions: F^{mid}, Q^{mid}; 4. Sample distribution and quantile: $F^{\sim}$, $Q^{\sim}$; 5. Comparison distribution and comparison density for continuous F, G; 6. Comparison distribution and comparison density for discrete F, G; 7. Information measures: Renyi, Chi-square; 8. Information for comparison density functions; 9. Information measures and entropy tests of fit; 10. Continuous versions of discrete distribution functions: F^c, Q^c; 11. Comparison distributions of one sample (continuous data); 12. One sample parameter estimation; 13. Compar-

Research supported by the U. S. Army Research Office

ison distributions of two samples (continuous data); 14. Two sample comparison density estimation.

2. Quantile domain functions: F, Q

The probability law of a random variable Y is described by its true ***distribution*** *function*

$$F(y) = \text{Prob}[Y \leq y], \quad -\infty < y < \infty,$$

and true ***quantile*** function

$$Q(u) = F^{-1}(u) = \inf[y : F(y) \geq u].$$

The most famous parametric model $F(y; \theta)$ is the Normal Distribution:

$$\begin{aligned} F(y; \mu, \sigma) &= \Phi((y - \mu)/\sigma), \\ \Phi(y) &= \int_{-\infty}^{y} \phi(x)dx, \\ \phi(x) &= (2\pi)^{-.5} \exp(-.5x^2), \\ Q(u; \mu, \sigma) &= \mu + \sigma\Phi^{-1}(u). \end{aligned}$$

Every random variable Y has a ***probability mass*** function, defined

$$p(y) = \text{Prob}[Y = y].$$

One can define $p(y)$ analytically in terms of the distribution function $F(y)$ as the jump in F at y. Similarly the ***spacing*** function of Y, denoted $sp(u)$, is defined as the jump in the quantile function Q at u; the jump at u is the difference between the right hand and left hand limits at u.

Continuous random variables obey $p(y) = 0$ for all y. Discrete random variables obey

$$\sum_{\text{all } y} p(y) = 1.$$

We call a random variable ***bi-continuous*** if both p and sp are identically zero. One of our goals is to unify data analysis for continuous and discrete data.

An (absolutely) continuous random variable Y has a distribution function which is determined by its probability density function $f(y) = F'(y)$. It is bi-continuous if $f(y) > 0$ for all y satisfying $0 < F(y) < 1$.

For a continuous random variable the ***density quantile*** function is defined by

$$fQ(u) = f(Q(u)).$$

Then $Q(u)$ has a ***quantile density*** $q(u) = Q'(u)$ satisfying

$$q(u) = 1/fQ(u), fQ(u)q(u) = 1.$$

To prove this important formula verify that the indefinite integral of $1/fQ(u)$ equals $Q(u)$, or differentiate the identity $FQ(u) = u$ which holds for continuous F.

Quantile data analysis implements methods of Data Analysis in the distribution F and the quantile Q domains. Parallel functions in the two domains are: F, Q; p, sp; f, q. Important interpretations are given by $\log f$ and $\log q$. The inverse of f is not used but the inverse of q is important and is given by fQ.

Three general properties of quantile functions are:

$F(Q(u)) = u$ if F is continuous at $y = Q(u)$;

$F(y) \geq u$ if and only if $Q(u) \leq y$;

$F^{-1}_{g(Y)}(u) = g(F^{-1}_Y(u))$ if g is a function with the mathematical properties of a quantile function: non-decreasing and left-continuous.

Two important applications of quantile functions are concerned with transforming Y to and from a random variable U which is Uniform[0,1]:

Y and $Q(U)$ are identically distributed (since $Q_{Q(U)}(u) = Q(Q_U(u)) = Q(u)$).

$F(Y)$ and U are identically distributed if F is continuous (since $Q_{F(Y)}(u) = F(Q(u)) = u$).

3. Mid-distribution and quantile functions: F^{mid}, Q^{mid}

We define several versions of the distribution function which we believe should play important roles in statistical data analysis. Versions of F and Q which we believe should be used routinely in the theory and practice of statistics are the mid-distribution function $F^{mid}(y)$, defined by

$$F^{mid}(y) = F(y) - .5p(y),$$

and the mid-quantile function defined by

$$Q^{mid}(u) = Q(u) + .5sp(u).$$

Note that $F(y)$ is right continuous, and $Q(u)$ is left continuous. Note that F^{mid} and Q^{mid} are not inverses of each other.

We recommend as the definition of the ***probability integral*** transformation (or ***rank transform***) of a continuous or discrete random variable Y

$$W = F^{mid}(Y);$$

it has mean $E[W] = .5$ and variance

$$\text{VAR}[Y] = (1/12)(1 - \sum_{\text{all } y} p^3(y)).$$

It is easy to verify this for a continuous Y (then W is Uniform[0,1]) and for a Bernoulli random variable Y taking values 0 and 1 with probabilities q and p; $F^{mid}(0) = .5q$ and $F^{mid}(1) = 1 - .5p$.

4. Sample distribution and quantile: $F^{\sim}$, $Q^{\sim}$

Our approach to data analysis of a data set $Y_1, \ldots, Y_n$ involves defining sample versions of F and Q. The initial way to represent a sample is to form its sample distribution function

$$F^{\sim}(y) = \text{ fraction of sample } \leq y$$

and its sample quantile function

$$Q^{\sim}(u) = F^{\sim -1}(u)$$

Explicit formulas for $F^{\sim}$ and $Q^{\sim}$ are expressed in terms of the distinct values in the sample, denoted $v_1 < \ldots < v_k$, their relative frequencies

$$p_j{}^{\sim} = \text{ fraction of sample } = v_j,$$

and their cumulative relative frequencies

$$u_j{}^{\sim} = p_1{}^{\sim} + \ldots + p_j{}^{\sim}, j = 1, \ldots, k.$$

Define $u_0{}^{\sim} = 0, v_0 = -\infty, v_{k+1} = \infty$.

The sample distribution function is discrete (piecewise constant) satisfying (for $j = 0, 1, \ldots, k$)

$$F^{\sim}(y) = u_j{}^{\sim} \text{ for } v_j \leq y < v_{j+1}.$$

The sample quantile function is discrete (piecewise constant) satisfying (for $j = 1, \ldots, k$)

$$Q^{\sim}(u) = v_j \text{ for } u_{j-1}{}^{\sim} < u \leq u_j{}^{\sim}.$$

We summarize these formulas by saying $F^{\sim}$ and $Q^{\sim}$ are piecewise constant between their values

$$F^{\sim}(v_j) = u_j{}^{\sim},$$
$$Q^{\sim}(u_j{}^{\sim}) = v_j.$$

5. Comparison distribution and comparison density for continuous F, G

Let F_θ be a specified continuous distribution (satisfying $Q_\theta F_\theta(y) = y$), which is a model for F, the unknown true distribution function of a continuous random variable Y. An important conceptual tool in statistical data analysis is transforming Y to the random variable $W = F_\theta(Y)$ which has quantile function

$$Q_W(u) = F_\theta(Q_Y(u))$$

and distribution function

$$F_W(y) = F_Y(Q_\theta(u))$$

How does one benefit by transforming probability law estimation problems to probability law estimation for a variable W on the unit interval? One could form

an estimator of the probability density of Y from an estimator of the probability density of W by sample analogies of the formulas relating their probability densities

$$f_W(y) = f_Y(Q_\theta(u))/f_0 Q_\theta(u),$$
$$f_Y(y) = f_0(y) f_W(F_\theta(y)).$$

Estimation of $f_W(u)$ provides estimation of $f_Y(y)$. One could form an estimator of the quantile and quantile density function of Y by sample analogues of the formulas

$$Q_Y(u) = Q_\theta(Q_W(u)),$$
$$q_Y(u) = q_0(Q_W(u)) q_W(u)$$

which require estimation of $Q_W(u)$ and $q_W(u)$. In our view the conceptual importance of the transformation to W comes from interpreting q_W as a comparison density function $d(u; F_Y, F_\theta)$.

To two distribution functions F and G we associate a comparison distribution function on $0 < u < 1$, denoted $D(u) = D(u; G, F)$. We consider three cases: both continuous; both discrete; one continuous and one discrete.

When F and G are continuous we define

$$D(u; G, F) = F(G^{-1}(u))$$

with comparison density function $d(u) = D'(u)$ given by

$$d(u) = d(u; G, F) = f(G^{-1}(u))/g(G^{-1}(u)).$$

We assume that $f(y) > 0$ implies $g(y) > 0$. Then $D(0) = 0$, $D(1) = 1$. We call F and G equivalent if $f(y) > 0$ if and only if $g(y) > 0$.

In terms of comparison distribution functions we express the quantile and distribution functions of $W = F_\theta(Y)$:

$$Q_W(u) = D(u; F_Y, F_\theta) = D(u;\ \text{data,model})$$
$$F_W(u) = D(u; F_\theta, F_Y) = D(u:\ \text{model,data}).$$

6. Comparison distribution and comparison density for discrete F, G.

When F and G are discrete we assume that their respective probability mass functions $p_F(y)$ and $p_G(y)$ satisfy

$$p_F(y) > 0 \text{ implies } p_G(y) > 0.$$

We call F and G equivalent if $p_F(y) > 0$ if and only if $p_G(y) > 0$. In the discrete case we define first the comparison density

$$d(u) = d(u; G, F) = p_F(G^{-1}(u))/p_G(G^{-1}(u))$$

and define its integral $D(u) = D(u; G, F)$ by

$$D(u) = D(u; G, F) = \int_0^u d(u')du'.$$

Our assumptions guarantee that $D(1) = 1$.

Analogues of this definition will be given in section 11 for F continuous and G discrete based on the following characterization of $D(u)$; it is a $P-P$ plot obtained by joining linearly the points

$$(0,0); (G(v_j), F(v_j)), j = 1, ..., k; (1,1)$$

where $v_1 < \ldots, v_k$ are the distinct values at which G jumps (which we have assumed to include all values at which F jumps).

We call our approach to data analysis "functional" because it emphasizes forming and smoothing functions on the interval $0 \leq u \leq 1$; raw estimators $d\hat{}(u; F_Y, F_\theta)$ and $d\hat{}(F_\theta, F_Y)$ are smoothed to form estimators $d\hat{}(u; F_Y, F_\theta)$ and $d\hat{}(u; F_\theta, F_Y)$. The graphs it provides for graphical data analysis are pictures of functions.

7. Information measures: Renyi, Chi-square

Comparison densities provide insight into information methods because information measures of univariate distributions can be expressed in terms of $d(u; F, G)$. Information measures play a central role in statistical data analysis because they provide tools to measure the "distance" between two probability distributions F and G. The (Kullback-Liebler) information divergence is defined (Kullback (1959)) by (our definitions differ from usual definitions by a factor of 2)

$$I(F;G) = (-2) \int_{-\infty}^{\infty} \log(g(x)/f(x)) f(x) dx$$

when F and G are continuous with probability density functions $f(x)$ and $g(x)$. When F and G are discrete, with probability mass functions $p_F(x)$ and $p_G(x)$, information divergence has an analogous definition:

$$I(F;G) = (-2) \sum \log\{p_G(x)/p_F(x)\} p_F(x).$$

An information decomposition of information divergence is

$$I(F;G) = H(F;G) - H(F),$$

in terms of entropy $H(F)$ and cross-entropy $H(F;G)$:

$$H(F) = (-2) \int_{-\infty}^{\infty} \{\log f(x)\} f(x) dx,$$

$$H(F;G) = (-2) \int_{-\infty}^{\infty} \{\log g(x)\} f(x) dx.$$

Adapting the fundamental work of Renyi (1961) we define Renyi information

of index λ. For continuous F and G: for $\lambda \neq 0, -1$

$$IR_\lambda(F;G) = \frac{2}{\lambda(1+\lambda)} \log \int \left\{ \frac{g(y)}{f(y)} \right\}^{1+\lambda} f(y)dy$$
$$= \frac{2}{\lambda(1+\lambda)} \log \int \left\{ \frac{g(y)}{f(y)} \right\}^{\lambda} g(y)dy$$
$$IR_0(F;G) = 2 \int \left\{ \frac{g(y)}{f(y)} \log \frac{g(y)}{f(y)} \right\} f(y)dy$$
$$= 2 \int \left\{ \frac{g(y)}{f(y)} \log \frac{g(y)}{f(y)} - \frac{g(y)}{f(y)} + 1 \right\} f(y)dy$$
$$IR_{-1}(F;G) = -2 \int \left\{ \log \frac{g(y)}{f(y)} \right\} f(y)dy$$
$$= -2 \int \left\{ \log \frac{g(y)}{f(y)} - \frac{g(y)}{f(y)} + 1 \right\} f(y)dy$$

An analogous definition holds for discrete F and G.

The second definition provides: (1) extensions to non-negative functions which are not densities, and also (2) a non-negative integrand which can provide diagnostic measures at each value of y.

Renyi information, for $-1 < \lambda < 0$, is equivalent to Bhattacharyya distance. Hellinger distance corresponds to $\lambda = -.5$.

In addition to Renyi information divergence (an extension of information statistics) one uses as information divergence between two non-negative functions an extension of chi-square statistics which has been developed by Read and Cressie (1988). For $\lambda \neq 0, -1$, Chi-square divergence of index λ is defined for continuous F and G by

$$C_\lambda(F;G) = \int B_\lambda \left(\frac{g(y)}{f(y)} \right) f(y)dy$$

where

$$B_\lambda(d) = \frac{2}{(1+\lambda)} \left\{ d \left(\frac{d^\lambda - 1}{\lambda} \right) - d + 1 \right\}$$
$$B_0(d) = 2\{d \log d - d + 1\}$$
$$B_{-1}(d) = -2\{\log d - d + 1\}$$

Important properties of $B_\lambda(d)$ are:

$$B_\lambda(d) \geq 0, B_\lambda(1) = B'_\lambda(1) = 0,$$
$$B'_\lambda(d) = \frac{2}{\lambda}\left(d^\lambda - 1\right), B''_\lambda(d) = 2d^{\lambda-1}$$
$$B_1(d) = (d-1)^2$$
$$B_0(d) = 2(d\log d - d + 1)$$
$$B_{-.5}(d) = 4\left(d^{.5} - 1\right)^2$$
$$B_{-1}(d) = -2(\log d - d + 1)$$
$$B_{-2}(d) = d\left(d^{-1} - 1\right)^2$$

An analogous definition holds for discrete F and G. Axiomatic derivations of information measures similar to C_λ are given by Jones and Byrne (1990).

The Renyi information and chi-square divergence measures are related:

$$IR_0(F;G) = C_0(F;G)$$
$$IR_{-1}(F;G) = C_{-1}(F;G)$$

For $\lambda \neq 0, -1$,

$$IR_\lambda(F;G) = \frac{2}{\lambda(1+\lambda)}\log\left\{1 + \left(\frac{\lambda(1+\lambda)}{2}\right)C_\lambda(F;G)\right\}$$

Interchange of F and G is provided by the *Lemma*: when F and G are equivalent,

$$C_\lambda(F;G) = C_{-(1+\lambda)}(G;F)$$
$$IR_\lambda(F;G) = IR_{-(1+\lambda)}(G;F)$$

Example: Renyi information divergence of two zero mean univariate normal distributions. Let P_j be the distribution on the real line corresponding to Normal $(0, K_j)$ with variance K_j. Let $\kappa = \frac{K_2}{K_1}$. Then

$$d(u;P_2,P_1) = \kappa^{.5}\exp\left\{-.5(\kappa-1)|\Phi^{-1}(u)|^2\right\}$$
$$IR_{-1}(P_2;P_1) = \kappa - 1 - \log\kappa,$$
$$IR_\lambda(P_2;P_1) = (1/\lambda)\{\log\kappa - (1+\lambda)^{-1}\log\{1 + (1+\lambda)(\kappa-1)\}_+\}$$
$$C_\lambda(P_2;P_1) = \{2/\lambda(1+\lambda)\}\,\kappa^{.5(1+\lambda)}\{1 + (1+\lambda)(\kappa-1)\}_+^{-.5}$$

8. Information for comparison density functions

Information divergence $I(F;G)$ is a concept that works for both multivariate and univariate distributions. In the univariate case we are able to relate $I(F;G)$ to the concept of comparison density $d(u;F,G)$,

For a density $d(u)$, $0 < u < 1$, Renyi information (of index λ), denoted $IR_\lambda(d)$, is non-negative and measures the divergence of $d(u)$ from uniform density $d_0(u) = 1$, $0 < u < 1$. It is defined:

$$IR_0(d) = 2\int_0^1 \{d(u)\log d(u)\}du = 2\int_0^1 \{d(u)\log d(u) - d(u) + 1\}\,du$$

$$IR_{-1}(d) = -2\int_0^1 \{\log d(u)\}du = -2\int_0^1 \{\log d(u) - d(u) + 1\}\,du$$

for $\lambda \neq 0$ or -1

$$IR_\lambda(d) = \{2/\lambda(1+\lambda)\}\log\int_0^1 \{d(u)\}^{1+\lambda}du$$

$$= \{2/\lambda\,(1+\lambda)\}\log\int_0^1 \Big(\{d(u)\}^{1+\lambda} - (1+\lambda)\,\{d(u)-1\}\Big)\,du.$$

To relate comparison density to information divergence we use the concept of Renyi information IR_λ which yields the important identity (and interpretation of $I(F;G)$!)

$$I(F;G) = (-2)\int_0^1 \log d(u;F,G)du$$

$$= IR_{-1}(d(u;F,G)) = IR_0(d(u;G,F)).$$

For a density $d(u)$, $0 < u < 1$, define

$$C_\lambda(d) = \int_0^1 B_\lambda(d(u))du.$$

The comparison density again unifies the continuous and discrete cases. One can show that for univariate F and G

$$C_\lambda(F,G) = C_\lambda(d(u;F,G))$$

For a random sample of a random variable with unknown probability density f, maximum likelihood estimators $\hat{\theta}$ of the parameters of a finite parameter model f_θ of the probability density f can be shown to be equivalent to minimizing

$$IR_{-1}(\hat{f};f_\theta) = IR_{-1}\left(d(u;\hat{F},F_\theta)\right)$$

where $\tilde{f}$ is a raw estimator of f (initially, a symbolic sample probability density formed from the sample distribution function $\tilde{F}$).

9. Information measures and entropy tests of fit

To test the goodness of fit hypothesis for a continuous random variable Y

$$H_0 : F_Y(y) = F_\theta(y),$$

many statistics have been proposed which start with the probability integral transformation

$$W = F_\theta(Y)$$

for which the goodnes of fit hypothesis is H_0: W is Uniform[0,1].

An entropy test of fit is Moran's statistic which transforms $Y_1, \ldots, Y_n$ to $W_1, \ldots, W_n$ and forms the order statistics $W(1;n) < \ldots < W(n;n)$ with spacings

$$S_i(\theta) = W(i;n) - W(i-1;n), j = 1, \ldots, n+1,$$

defining $W(0,n) = 0$, $W(n+1;n) = 1$ Moran's statistic is often defined as

$$\sum_{i=1}^{n+1} -\log S_i(\theta),$$

We prefer to define it as

$$\tilde{M}(\theta) = (n+1)^{-1} \sum_{i=1}^{n+1} (-2) \log(n+1) S_i(\theta).$$

Under the null hypothesis, it is asymptotically normal with mean 2γ ($\gamma = .57722$, Euler's constant), and variance

$$4(n+1)^{-1}((\pi^2/6) - 1).$$

Small sample asymptotic chi-square and beta distributions (given by Smethurst and Mudholkar (1991)) *are more appropriate for an entropy interpretation.*

In order to understand and extend $\tilde{M}(\theta)$, we regard it as an estimator of a quantity $M(\theta)$ defined by probability theory. The original observation Y has true distribution function F_Y and quantile function $Q_Y(u) = F_Y^{-1}(u)$. The transformation $W = F_\theta(Y)$ has quantile function $Q_W(u) = F_\theta(Q_Y(u))$ and distribution function $F_W(u) = F_Y(Q_\theta(u))$; W has quantile density $q_W(u) = f_\theta(Q_Y(u))/f_Y Q_Y(u)$, and entropy

$$H(W) = \int_0^1 (-2 \log f_W(w)) f_W(w) dw = \int_0^1 2 \log q_W(u) du.$$

Under the null hypothesis $H_0 : F_Y = F_\theta$, $q_W(u)$ is identically 1, and $H(W) = 0$.

Note $-H(W)$, the neg-entropy of W, is non-negative and is the population parameter, denoted $M(\theta)$, which is being non-parametrically estimated by Moran's statistic $\tilde{M}(\theta)$.

How do we benefit from estimating entropy (or neg-entropy) of W? It provides tests of H_0 and can provide (through suitable analogues of Akaike Information Criterion) insight about selection of alternative models to fit when one rejects H_0. Thus understanding and improving Moran's statistic requires us to solve problems of density estimation, especially estimation of the smooth comparison density $d(u; F_Y, F_\theta)$ from raw estimators

$$\tilde{d}(u; F_Y, F_\theta) = \tilde{d}(u; \tilde{F}^c, F_\theta),$$

where $\tilde{F}^c$ is a continuous version of the discrete distribution function $\tilde{F}$.

Another important interpretation of $M(\theta)$ is $M(\theta) = I(f; f_\theta)$, the Kullback information divergence between the true $F(y)$ and the model $F_\theta(y)$.

10. Continuous versions of discrete distribution functions

To compare a continuous and a discrete distribution we propose forming a continuous distribution function version of a discrete one. To estimate a continuous distribution function F from data we recommend first forming our continuous version $\tilde{F}^c$ of the discrete distribution function given by the sample distribution function $\tilde{F}$ formed from the data.

For discrete data we recommend estimating the continuous version F^c of its discrete distribution function F. We conjecture that these recommendations provide a unified theory of discrete and continuous data analysis as well as improved methods of continuous data analysis.

To define the continuous version of a discrete distribution F, we assume that it can be described by a finite number of points (v_j, u_j) such that

$$F(v_j) = u_j \text{ for } j = 1, \ldots, k.$$

Note that $u_k = 1$. Define $u_0 = 0$; then $0 = u_0 < u_1 < \ldots < u_k = 1$.

Its quantile function $Q(u)$ is discrete and satisfies, for $j = 1, \ldots, k$,

$$Q(u_j) = v_j.$$

Define "mid-values" v_j^c, $j = 0, \ldots, k$, by

$$v_0^c = v_1, v_k^c = v_k,$$
$$v_j^c = .5(v_j + v_{j+1}) \text{ for } j = 1, \ldots, k-1.$$

Define F^c and Q^c to be piecewise linear between its values (for $j = 0, 1, \ldots, k$)

$$Q^c(u_j) = v_j^c,$$
$$F^c(v_j^c) = u_j$$

We call F^c and Q^c continuous (piecewise linear) versions of the discrete (piecewise constant) functions F and Q.

It is interesting to compare the continuous version of a discrete distribution to its mid-distribution F^{mid} whose definition we recall; define $p_j = u_j - u_{j-1}$ and

$$F^{mid}(v_j) = u_j - .5p_j \text{ for } j = 1, ..., k.$$

One conjectures that approximately (and exactly when v_j are equi-spaced)

$$F^c(v_j) = u_j - .5p_j,$$

so that approximations to F^c are also approximations to F^{mid}.

To justify our view that these concepts are very natural, we would argue that the continuity correction when one approximates a discrete distribution by a continuous one (say the binomial by the normal) can be explained by regarding the limiting continuous distribution as approximating F^c and F^{mid} rather than F.

Let F be the distribution function of a Binomial(n, p) random variable. The continuity correction says that (for $x = 0, 1, \ldots, n$) approximately

$$F(x) = F^c(x + .5) = \Phi((x + .5 - np)/(np(1-p))^{.5})$$

Note that (for $x = 0, 1, \ldots, n$) approximately

$$F^c(x) = F^{mid}(x) = \Phi((x - np)/(np(1-p)^{.5}).$$

11. Comparison distributions of one sample (continuous data)

Given a sample $Y_1, \ldots, Y_n$ from a continuous distribution F, we recommend as the first step in data analysis to compute and plot F^{-c} and Q^{-c}, the continuous versions of the discrete sample distribution and quantile functions. We regard Q^{-c} as a raw estimator of the true quantile Q which provides a minimal amount of smoothing of the observations.

The process of fitting a model to the data can be formulated in terms of a specified continuous distribution function F_θ (whose form may be guessed from a visual examination of Q^{-c}, normalized at $u = .5$ to equal 0 and to have slope 1 (compare Parzen (1986))). We not only estimate F but also test the goodness of fit hypothesis $H_0 : F = F_\theta$. To motivate our approach let us review some methods of testing a goodness of fit hypothesis H_0 for continuous data.

A graphical diagnostic of H_0 is the $Q - Q$ plot which looks for linearity in the graph of

$$(Q_\theta(F^{-mid}(v_j)), v_j), j = 1, \ldots, k;$$

an alternative is the $Q - Q^c$ plot which graphs the points

$$(Q_\theta(u_j), v_j^c) = (Q_\theta(F^{-}(v_j)), v_j^c), j = 1, ..., k-1.$$

Goodness of fit tests of H_0 have traditionally been expressed by statisticians as measures of $F^{\sim}(y) - F_\theta(y)$, the difference of distribution functions. Typical measures are non-linear functionals such as

$$D[\text{KolmogorovSmirnov}] = \text{supremum}_y|F^{\sim}(y) - F_\theta(y)|$$

$$D[\text{CramervonMises}] = \int_{-\infty}^{\infty} |F^{\sim}(y) - F_\theta(y)|^2 dF_\theta(y).$$

Goodness of fit compares probabilities; we believe that probabilities $p^{\sim}$ and $p^{\wedge}$, representing *data and model, should be compared not by their differences but by their ratio!* In symbols, measure $(p^{\sim}/p^{\wedge}) - 1$ rather than $p^{\sim} - p^{\wedge}$. Therefore goodness of fit tests should be based on measures of the difference from the identity function $D_0(u) = u, 0 < u < 1$, of comparison distribution functions. Goodness of fit tests for uniformity are traditionally based on

$$D(u; F_\theta, F^{\sim}) = F_{W^{\sim}}(u) = F^{\sim}(Q_\theta(u)), 0 < u < 1,$$

the sample distribution function of the probability integral transformation $W = F_\theta(Y)$. Traditional maximum likelihood estimators of θ are chosen by the criterion that the sample quantile function

$$D(u; F^{\sim}, F_\theta) = Q_{W^{\sim}}(u) = F_\theta(Q^{\sim}(u), 0 < u < 1,$$

has minimum Kullback information distance from $D_o(u) = u$.

This paper proposes that we need to overcome the problem that $F^{\sim}$ and $Q^{\sim}$ are discrete and are not directly covered by our definitions of comparison distribution functions; we recommend that data analysis be based on the definitions below of continuous raw comparison functions, denoted $D^c(u; F^{\sim}, F_\theta)$ and $D^c(u; F_\theta, F^{\sim})$. Analogously one defines comparison distribution functions, denoted $D^c(u; G, F)$ and $D^c(u; F, G)$ rather than $D(u; G, F)$ and $D(u; F, G)$, when F is continuous and G is discrete.

Recall $F^{\sim}(v_j) = u_j$ for $j = 1, \ldots, k$. Let v_j^c, $j = 0, \ldots, k$ be mid values. Define for $j = 1, \ldots, k-1$

$$w_j = F_\theta(Q^{\sim c}(u_j)) = F_\theta(v_j^c)).$$

Assume $0 < w_1 < \ldots w_{k-1} < 1$.

Define $Q_W^c(u)$ as a piecewise linear curve connecting the values

$$(0,0), (u_j, w_j) \text{ for } j = 1, ..., k-1, (1,1).$$

Define $F_W^c(u)$ as a piecewise linear curve connecting the values

$$(0,0), (w_j, u_j) for j = 1, ..., k-1, (1,1).$$

The derivatives, denoted $q_W^c(u)$ and $f_W^c(u)$ respectively, are sample quantile density and probability density functions. Define $d^c(u; F^{\sim}, F_\theta) = q_W^c(u), d^c(u; F_\theta, F^{\sim}) = f_W^c(u)$.

One smooths raw densities to form smooth estimators, denoted

$$q_W\hat{}(u), Q_W\hat{}(u), f_W\hat{}(u), F_W\hat{}(u).$$

The adequacy of the smoothing can be judged by comparing on one graph Q_W^c and $Q_W\hat{}$, and comparing on one graph F_W^c and $F_W\hat{}$. In this way one can develop $f_Y\hat{}(y)$ and $Q_Y\hat{}(u)$.

12. One sample parameter estimation.

Regular *maximum likelihood* estimators $\theta\hat{}$ are parameter values minimizing

$$\int_0^1 -\log f_\theta(Q\tilde{}(u))du$$

or equivalently minimizing the negative of the average log likelihood

$$-L(\theta) = (1/n)\sum_{j=1}^{n} -\log f_\theta(Y(j;n))$$

A *maximum spacings* estimator, denoted $\theta\hat{}$, can be obtained (compare Ranneby (1984)) by minimizing (with respect to all possible parameter values θ) the neg-entropy

$$2\int_0^1 -\log d^c(u; F\tilde{}, F_\theta)du$$

or equivalently minimizing

$$-2\sum_{j=1}^{k}(u_j - u_{j-1})\log\left(\left(F_\theta(Q\tilde{}^c(u_j)) - F_\theta(Q\tilde{}^c(u_{j-1}))/(u_j - u_{j-1})\right)\right.\,.$$

In this expression, logarithm is taken after integration rather than before; consequently it provides estimators in non-regular cases.

Minimum information estimators (more precisely, minimum Renyi information of index λ estimators) $\theta\hat{}$, minimize (for $-1 < \lambda < 0$, and especially $\lambda = -.5$ (compare Beran (1977))).

$$IR_\lambda\left(d^c\left(u; F\tilde{}, F_\theta\right)\right).$$

Regular maximum likelihood estimators (which correspond to $\lambda = -1$) satisfy the estimating equations

$$\int_0^1 S_{\theta_i}\left(Q\tilde{}^c(u)\right)du = 0$$

where

$$S_{\theta_i}(y) = \frac{\partial}{\partial\theta_i}\log f_\theta(y)$$

is the score function of component θ_i of the vector parameter θ. Minimum information estimators satisfy the estimating equations

$$\int_0^1 (d^c(u; F^{\sim}, F_\theta))^{1+\lambda} S_{\theta_i}(Q^{\sim c}(u))\, du = 0.$$

Minimum information estimators provide *robust* estimators.

To test if robust minimum information estimators of a given data set are to be preferred to regular maximum likelihood estimators, one could test if the latter satisfy the estimating equations of the former. The theory and practice of this "test for robustness" are open research problems.

13. Comparison distributions of two samples (continuous data)

A central problem of statistics is test

$$H_0 : F_1 = F_2,$$

the equality of two continuous distribution functions F_1 and F_2. The data are assumed to be independent observations of a first sample

$X_1, \ldots, X_{n_1}$ assumed to be distributed as F_1,

and a second sample

$Y_1, \ldots, Y_{n_2}$ assumed to be distributed as F_2.

Let $F_1^{\sim}$ and $F_2^{\sim}$ denote the sample distribution functions of the two samples.

The pooled sample consists of all X and Y values; its sample distribution function is denoted $F^{\sim}$. Let $n = n_1 + n_2$ be the total sample size. Let $p_j = n_j/n$ be the fraction of the pooled sample in the j-th sample. One can represent

$$F^{\sim} = p_1 F_1^{\sim} + p_2 F_2^{\sim}.$$

it is an estimator of the true pooled distribution

$$F = p_1 F_1 + p_2 F_2.$$

The novelty of our approach to testing H_0 is our proposed comparison distribution function

$$D^{\sim}(u) = D(u; F^{\sim}, F_1^{\sim})$$

which estimates $D(u) = D(u; F, F_1)$. Because $F^{\sim}$ and $F_1^{\sim}$ are both discrete, the comparison distribution $D^{\sim}(u)$ is defined in terms of the comparison density function

$$d^{\sim}(u) = d(u; F^{\sim}, F_1^{\sim}).$$

The asymptotic distribution of $D^{\sim}(u)$ as an estimator of $D(u) = D(u; F, F_1)$ can be shown to be the same as the Pyke-Shorack two sample process. The mathematical statistics is the same, but the data analysis is greatly improved.

Linear rank statistics to test H_0 can be represented as linear functionals

$$< J(u), d^{\sim}(u) >= \int_0^1 J(u) d^{\sim}(u) du$$

for suitable score functions $J(u)$. The Wilcoxon statistic (which tests for a shift in location) corresponds to $J(u) = u$, whose orthonormal version is $J(u) = 12^{.5}(u - .5)$.

14. Two sample comparison density estimation.

We propose that the best summary of two sample data analysis is not a p value of a linear rank statistic but a smooth density estimator $\hat{d}(u)$ of the true comparison density $d(u) = d(u; F, F_1)$.

The asymptotic distribution of estimators $\hat{d}(u)$ which smooth $\hat{d}(u)$ is best understood by normalizing it to be between 0 and 1 by defining

$$\tilde{p}(u) = p_1 \hat{d}(u)$$

whose smooth estimators are denoted $\hat{p}(u)$. Their asymptotic distribution theory (outlined below) is developed on the assumption that they are estimators of a true comparison density $d(u) = d(u; F, F_1)$; let $p(u) = p_1 d(u)$. The asymptotic variance of $\hat{p}(u)$ can be shown to be proportional to $p(u)(1 - p(u))$ which means that its asymptotic distribution is similar to that of an estimator $\hat{p}$ of a probabiity p.

In contrast the asymptotic variance in the one sample case of the smooth quantile density estimator $\hat{q}$ is proportional to q^2, and the asymptotic variance of the smooth probability density estimator $\hat{f}$ is proportional to f if $\hat{f}$ is a standard kernel estimator, and is proportional to f^2 if $\hat{f}$ is a nearest neighbor estimator.

One of the joys of a unifed framework for one sample and two sample data analysis is that it can comprehend and explain the different qualitative behavior of estimators of different types of densities. Parzen (1983) states, and outlines proofs of the following results about comparison density estimation.

A *kernel* comparison density estimator has the form

$$\hat{d}(u) = \int_0^1 K_M(u - t)\hat{d}(t)dt$$

where

$$K_M(t) = \sum_{v=-\infty}^{\infty} e^{-2\pi i t v} k_M(v)$$

$$k_M(v) = k\left(\frac{v}{M}\right)$$

We call M truncation point (or effective number of parameters); it is chosen as a function of n and tends to ∞, as n tends to ∞, at a suitable rate. Let

$$\overline{K^2} = \int_{-\infty}^{\infty} K^2(t)dt = \int_{-\infty}^{\infty} k^2(x)dx$$

One can show that (by letting M tend to ∞ at a suitable rate)

$$\lim_{n\to\infty} \frac{n}{M} \operatorname{Var}[\hat{p}(u)] = \overline{K^2} p(u)(1 - p(u))$$

The *numerical derivative* density estimator

$$d\hat{}(u) = (2h)^{-1}(D\tilde{}(u+h) - D\tilde{}(u-h)$$

corresponds to $M = 1/h$ and $k(x) = (\sin 2\pi x)/2\pi x$; it has asymptotic variance

$$\lim_{n\to\infty} hn \operatorname{Var}[p\hat{}(u)] = .5p(u)(1-p(u)).$$

Evaluate $\overline{K^2}$ from $K(t) = .5$ for $|t| \leq 1, 0$ otherwise.

An *autoregressive* (of order m) estimator has asymptotic variance

$$\lim_{n\to\infty} \frac{n}{m} \operatorname{Var}[p\hat{}(u)] = 2p(u)(1-p(u))$$

Model order selection techniques can be developed by adapting Akaike (1973), Atilgan and Bozdogan (1990), Sakamoto, Ishiguro, Kitagawa (1983).

To obtain one sample probability density estimator results from the two sample results, let the first sample have unknown distribution F_1, the second sample have known distribution F_2, and let n_2 be very large. Then the pooled distribution $F = F_2$, $d(u)$ equals $d(u; F_2, F_1)$, and p_1 tends to 0. The kernel probability density estimator of f_1 has asymptotic variance

$$\lim_{n_1\to\infty} \frac{n_1}{M} \operatorname{Var}[d\hat{}(u; F_2, F_1)] = \overline{K^2} d(u; F_2, F_1)$$

since

$$p(u) = p_1 f_1(F^{-1}(u))/f(F^{-1}(u)) = p_1 d(u; F_2, F_1).$$

References

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* Proc. of the Second International Symposium on Information Theory, B. N. Petrov and F. Csaki, Akademiai Kiado, Budapest, pp. 267-281. Reprinted in *Breakthroughs in Statistics*, I. edited by S. Kotz and N. L. Johnson, Springer-Verlag: New York, 1992.

Arimoto, S. (1971). "Informational-theoretical considerations on estimation problems," *Inf. and Control*, 19, 181-194.

Atilgan, T. and Bozdogan, H. (1990). "Selecting number knots fitting cardinal *B*-splines for density estimation using AIC." *J. Japan Statist. Soc*, 20 (2), 179-190.

Beran, R. J. (1977). "Minimum Hellinger Distance Estimates for Parametric Models." *Annals of Statistics*, 5, 445-463.

Jones, L. K. and Byrne, C. L. (1990). "General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis," *IEEE Transactions on Information Theory*, vol. 36, no. 1, pp. 23-30.

Kullback, S. (1959). *Information Theory and Statistics.* New York: Wiley.

Parzen, E. (1979). "Nonparametric Statistical Data Modeling", *Journal of the American Statistical Association*, (with discussion), 74. 105-131.

Parzen, E. (1983). "FUN.STAT Quantile Approach to Two Sample Statistical Data Analysis." Technical Report.

Parzen, E. (1986). "Quantile Spectral Analysis and Long Memory Time Series," *Journal of Applied Probability*, Vol. 23A, 41-55.

Ranneby, B. (1984). "The maximum spacing method," *Scand. J. Statist.*, 11, 93-112.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness of Fit Statistics for Discrete Multivariate Data*, Springer Verlag, New York.

Renyi, A. (1961). "On measures of entropy and information." *Proc. 4th Berkeley Symp. Math. Statist. Probability, 1960*, 1, 547-561. University of California Press: Berkeley.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1983). *Akaike Information Criterion Statistics*, D. Reidel: Boston.

Smethurst, P. A. and Mudholkar, G. S. (1991). "Two Facets of the Moran Statistics." *J. Statist. Comput. Simul.* 39, 215-220.